N (or n) vs. N - 1(or n - 1) re-visited

Thomas R. Knapp

Prologue

Over 40 years ago I (Knapp,1970) wrote an article regarding when you should use N and when you should use N - 1 in the denominators of various formulas for the variance, the standard deviation, and the Pearson product-moment correlation coefficient. I ended my "pro N" article with this sentence: "Nobody ever gets an average by dividing by one less than the number of observations." (page 626).

There immediately followed two other comments (Landrum, 1971; Games, 1971) concerning the matter of N vs. N - 1. Things were relatively quiet for the next few years, but the controversy has erupted several times since, culminating in a clever piece by Speed (2012) who offered a cash prize [not yet awarded] to the person who could determine the very first time that a discussion was held on the topic.

The problem

Imagine that you are teaching an introductory ("non-calculus") course in statistics. [That shouldn't be too hard. Some of you who are reading this might be doing that or have done that.] You would like to provide your students with their first formulas for the variance and for the standard deviation. Do you put N, N -1, n, or n - 1 in the denominators? Why?

Some considerations

1. Will your first example (I hope you'll give them an example!) be a set of data (real or artificial) for a population (no matter what its size)? I hope so.

N is fine, and is really the only defensible choice of the four possibilities. You never subtract 1 from the number of observations in a population; and nobody uses n to denote the population size.

2. Will that first example be for a sample?

N would be OK, if you always use N for sample size and use something like N_{pop} for population size. [Yes, I have seen N_{pop}.]

N -1 would be OK for the sample variance, if you always use N for sample size, you have a random sample, and you would like to get an unbiased estimate of the population variance; but it's not OK for the sample standard deviation. (The

square root of an unbiased estimate of a parameter is not an unbiased estimate of the square root of the parameter. Do you follow that?)

n would be OK for both the sample variance and the sample standard deviation, and is my own personal preference.

n - 1 would be OK for the sample variance, if you always use n for sample size, you have a random sample, and you would like to get an unbiased estimate of the population variance; but it's not OK for the sample standard deviation (for the same reason indicated for N - 1).

3. What do most people do?

I haven't carried out an extensive survey, but my impression is that many authors of statistics textbooks and many people who have websites for the teaching of statistics use a sample for a first example, don't say whether or not the sample is a random sample, and use n - 1 in the denominator of the formula for the variance and in the denominator of the formula for the standard deviation.

4. Does it really matter?

From a practical standpoint, if the number of observations is very large, no. But from a conceptual standpoint, you bet it does, no matter what the size of N or n. In the remainder of this paper I will try to explain why; identify the principal culprits; and recommend what we should all do about it.

Why it matters conceptually

A variance is a measure of the amount of spread around the arithmetic mean of a frequency distribution, albeit in the wrong units. My favorite example is a distribution of the number of eggs sold by a super market in a given month. No matter whether you have a population or a sample, or whether you use in the denominator the number of observations or one less than the number of observations, the answer comes out in "squared eggs". In order to get back to the original units (eggs) you must "unsquare" by taking the square root of the variance, which is equal to the standard deviation.

A variance is a special kind of mean. It is the mean of the squared differences (deviations) from the mean. A standard deviation is the square root of the mean of the squared differences from the mean, and is sometimes called "the root mean square".

The culprits

In my opinion, there are two sets of culprits. The first set consists of some textbook authors and some people who have websites for the teaching of

statistics who favor N - 1 (or n - 1) for various reasons (perhaps they want their students to get accustomed to n - 1 right away because they'll be using that in their calculations to get unbiased estimates of the population variance, e.g., in ANOVA) or they just don't think things through.

The second set consists of two subsets. Subset A comprises the people who write the software and the manuals for handheld calculators. I have an old TI-60 calculator that has two keys for calculating a standard deviation. One of the keys is labelled σ_n and the other is labelled σ_{n-1}. The guidebook calls the first "the population deviation"; it calls the second "the sample deviation" (page 5-6). It's nice that the user has the choice, but the notation is not appropriate. Greek letters are almost always reserved for population parameters, and as indicated above you don't calculate a population standard deviation by having in the denominator one less than the number of observations. Subset B comprises the people who write the software and the manuals for computer packages such as Excel, Minitab, SPSS, and SAS. All four of those use n - 1 as the default. [Good luck in trying to get the calculation using n.].

n + 1 [not the magazine]

Believe it or not, there are a few people who recommend the use of n + 1 in the denominator, because that produces the minimum mean squared error in estimating a population variance. See, for example, Biau & Yatracos (2012).

Degrees of freedom

Is it really necessary to get into degrees of freedom when first introducing the variance and the standard deviation? I don't think so. It's a strange concept (as Walker, 1940, pointed out many years ago) that students always have trouble with, no matter how you explain it. The number of unconstrained pieces of data? Something you need to know in order to use certain tables in the backs of statistics textbooks? Whatever.

Pearson r

For people who use n in the denominator for the sample variance and the sample standard deviation, the transition to the Pearson product-moment correlation coefficient is easy. Although there are at least 13 different formulas for the Pearson r (Rodgers & Nicewander, 1988; I've added a few more), the simplest to understand is $\sum z_X z_Y / n$, where the z's are the standard scores for the two variables X and Y that are to be correlated. The people who favor n - 1 for the standard deviation, and use that standard deviation for the calculation of the z scores, need to follow through with n - 1 in the denominator of the formula for Pearson r. But that ruins "the average cross-product of standard scores" interpretation. If they don't follow through with n - 1, they're just plain wrong.

A call to action

If you happen to be asked to serve as a reviewer of a manuscript for possible publication as an introductory statistics textbook, please insist that the authors provide a careful explanation for whatever they choose to use in the denominators for their formulas for the variance, the standard deviation, and the Pearson r. And if you have any influence over the people who write the software and the manuals for computer packages that calculate those expressions, please ask them to do the same. [I have no such influence. I tried very hard a few years ago to get the people at SPSS to take out "observed power" from some of its ANOVA routines. They refused to do so.]

References

Biau, G., & Yatracos, Y.G. (2012). On the shrinkage estimation of variance and Pitman closeness criterion. Journal de las Societe Francaise de Statistique, 153, 5-21. [Don't worry; it's in English.]

Games, P.A. (1971). Further comments on "N vs. N - 1". American Educational Research Journal, 8, 582-584.

Knapp, T.R. (1970). N vs. N - 1. American Educational Research Journal, 7, 625-626.

Landrum, W.L. (1971). A second comment on N vs. N - 1. American Educational Research Journal, 8, 581.

Rodgers, J.L., & Nicewander, W.A. (1988). Thirteen ways to look at the correlation coefficient. The American Statistician, 42, 59-66.

Speed, T. (December 19, 2012). Terence's stuff: n vs. n - 1. IMS Bulletin Online.

Walker, H.M. (1940). Degrees of freedom. Journal of Educational Psychology, 31, 253-269.